Identification and Characterization of Newsworthy Verbs in World News

Benjamin Nye
University of Pennsylvania
bepnye@gmail.com

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Abstract

We present a data-driven technique for acquiring domain-level importance of verbs from the analysis of abstract/article pairs of world news articles. We show that existing lexical resources capture some the semantic characteristics for important words in the domain. We develop a novel characterization of the association between verbs and personal story narratives, which is descriptive of verbs avoided in summaries for this domain.

1 Introduction

Summarization, either by people or machine, calls for the ability to identify important content. Computational approaches to identifying important content fall into the two extremes of a possible spectrum. On one end, the types of important information for a given domain and topic are predefined as information extraction templates defined by experts, as in the earliest approaches to multi-document summarization (Radev and McKeown, 1998) and the recently introduced guided summarization (Owczarzak and Dang, 2011). On the other extreme, traditional systems work only with indicators of importance coming solely from the input to be summarized, or possibly also from the context of the input, i.e . analyzing the anchor text of links to a webpage, or comments on a blog post or citations to a scientific article (Nenkova and McKeown, 2012).

Here we explore the feasibility of data-driven identification of important information in the world news domain. We specifically focus on the analysis of verbs, which is the first step of identifying event types of special interest. The goal is to collect evidence of verb importance globally, without regard to a particular input or its context. Such ideas have been explored in the past as subcomponents of extractive summarizers (Schiffman et al., 2002; Hong and Nenkova, 2014) or as features derived from small datasets for sentence compression (Woodsend and Lapata, 2012). In contrast, in our work we rely on large corpora and exclusively focus on the task of acquiring input independent indicators of importance. We also constrain our analysis to a single domain, which allows us to examine the semantic aspects of the verbs that may contribute to their perceived importance.

We leverage a dataset of human-written summaries of news articles to objectively ground the definition of word importance. Summaries are intended to convey important information while omitting the less important pieces, so words that are important in a newsworthy sense will occur more frequently in summaries. The same data and intuition was used recently to develop a large corpus for determining entity salience (Dunietz and Gillick, 2014).

We derive a list of over one thousand verbs that have statistically significant bias to appear in the summaries (important verbs) and verbs with higher rate of occurrence in the original articles (unimportant). This resource of verbs and their domain-level importance may be fruitfully exploited in models of summarization that do not use pre-defined templates but are richer than approaches that rely solely on analysis of the article text.

We furthermore seek to characterize the properties of words that are biased to occur more often

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1440–1445,
Denver, Colorado, May 31 – June 5, 2015.

in either summaries or in articles. We noticed that verbs that tended to be dis-preferred in the summaries related to personal narratives, in which people are described as private entities rather than public personas. We applied the same measures that we used to analyze domain-level importance in world news to a collection of labeled personal and nonpersonal blog entries. Characterizing verbs on the personal vs. nonpersonal dimension indeed turned out to beneficial for explaining domain-level importance of verbs in world news: personal narratives are not considered important in this domain and verbs that tended to get excluded from summaries also tended to appear more frequently in personal blog entries. This characterization offered broad coverage of the article vocabulary and additional explanatory power compared to a characterization derived from General Inquirer categories.[1]

The derived lexical resources may serve as shallow semantics for a range of language processing tasks such as summarization, news filtering and search.[2]

2 Determining domain-level importance

To determine domain-level importance, we use summaries and articles from the New York Times Annotated Corpus[3], a collection of NYT articles that includes genre tags and summaries written by library scientists. We use articles published in the world news section between 1996 and 2005 for a total of 36,69 article-summary pairs.

All of the documents were parsed with the Stanford Parser to obtain lemmatized forms of the words and part of speech tags. There are 2,634,850 tokens in the summaries and 32,587,740 tokens in the respective articles.

The overall verb frequency is very similar in the summaries (14.5%) and the articles (14.6%). In our analysis we reduce the corpus of summaries and original articles to only the verbs that occur in them. Then we compare the rate of occurrence of each verb in the two types of writing. Verbs used proportionally more in summaries are likely to correspond to events that are important, while verbs that occur more frequently in the articles are less likely to be related to the key topics of an article. To generate two classes of verbs representing important and non-important verbs, we consider two measurements: the difference of a verb's usage frequency between summaries and articles, and the statistical significance of this bias.

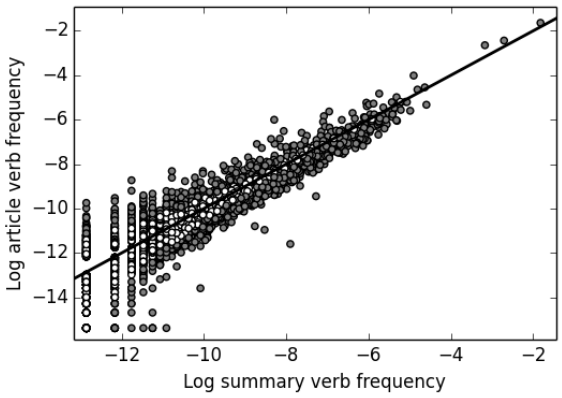

Figure 1: Word frequency in summaries vs. articles. The black line is where the frequencies are equal, and words plotted in grey have statistically significant differences in frequencies for the two classes.

Figure 1 shows the plot of each verb's probability in the summaries, $P_s(w_i)$ vs. in the articles, $P_a(w_i)$. Points above the line are verbs that occur more frequently in articles, while points below the line are more frequent in summaries. Dividing the points along this line produces two classes of verbs. We can further quantify how strongly a word is associated with its class using a variety of metrics (Monroe et al., 2008).

For this application, we chose to use the log odds ratio. To measure how much more likely a word w is to occur in a class c, we compute the odds of a word occurring in corpora type c, i.e summary (s) or original article (a):

$$Odds(w, c) = \frac{P(w \mid class = c)}{1 - P(w \mid class = c)}$$

The ratio of the odds with respect to the two different corpora is a measure of how much more frequently

[1]The two lists characterizing the domain-level importance and the personal–public dimension are available for download at `http://www.cis.upenn.edu/~nlp/software/importance.html`.

[2]Personal perspective verbs may not be important in reporting world news but may be excellent indicators of celebrity/gossip search for example.

[3]https://catalog.ldc.upenn.edu/LDC2008T19

a word is used in each case. To make the measure interpretable, we take the log of the odds ratio producing the final weight for a word:

$$\log\left(\frac{Odds(w,s)}{Odds(w,a)}\right) \simeq \log\left(\frac{P(w \mid class = s)}{P(w \mid class = a)}\right)$$

This metric gives an intuitive measure of the usage rate of words. For example, if a word occurs 3 times more often in the summaries it will be given a weight of $\log(3)$, and if it occurs three times more often in the articles it will be given a weight of $-\log(3)$.

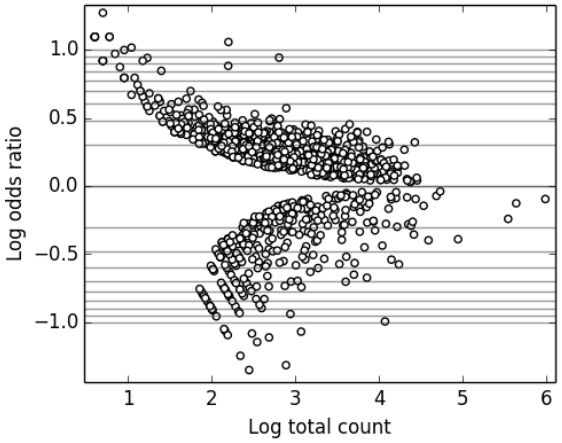

Figure 2: Frequency vs. log odds ratio for each word. Positive odds ratios correspond to summary-biased words and negative correspond to article-biased. The lines indicate integer usage ratios (1:3, 1:2, 1:1, 2:1, 3:1, etc).

However, this metric is unreliable for verbs with low counts in either class of texts. If a verb occurs five times in summaries and only once in articles, it is difficult to say if there is true signal for importance. As the number of counts of a verb in the two classes increases, so does our certainty about the significance of any observed differences in usage rates. To obtain a measure of the statistical significance of the domain-level importance weight of a verb, we treat the set of observed tokens as a Bernoulli trial where each token occurs either in a summary (success) or in an article (failure). We apply a binomial test to compute the probability of the observed distribution of tokens in the two types of text under the null hypothesis that the word has equal frequency in summaries and articles. The p-value from the test gives a measure of the certainty that a word is important and not. We can filter out words with p-values above a certain threshold. Moving the threshold closer to 0 enforces more and more certainty about the classification, reducing vocabulary size but also decreasing noise. Discarding verbs with a p-value of less than 0.05 reduces the vocabulary size from 3,924 words to 1,210.

Summary-biased
spur hail allege avert slay exile claim intensify extradite oust overturn underscore cite devastate weigh defuse injure curb defy resign suspect warn quell kidnap stir plot widen charge thwart revive
Article-biased
chant talk sleep hate graduate realize dress understand quote sound add drink sing refer read think imagine remember shout sit happen cry wave like thank love smile accord reply misstate

Table 1: Words with highest weights, drawn from verbs with frequency greater than the median verb frequency.

In Figure 2, the log odds ratio is plotted against the overall frequency for each verb after discarding unreliable verbs. The most extreme weights occur mostly for the infrequent words, even after filtering out low p-value words. Although these words have extremely high bias weights, they tend to be uncommon and not particularly informative. Examples include verbs such as "hostage-take", "muckrake", and "blaspheme." To get a clearer picture of trends in the verbs in each category, we show in Table 1 the verbs with the 30 highest and 30 lowest weights among the verbs in the 50th percentile of total counts across all documents.

In the following two sections, we turn to analyzing *why* certain verbs may be more important in the domain than others. First we examine the relationship between the summary- or article-bias of a word and categories in the General Inquirer lexicon. Then we develop a new characterization of verbs showing their association with person-centered perspective of the narrative.

3 General Inquirer

The General Inquirer lexicon provides a list of words manually annotated with a variety of tags (Stone et al., 1966). We considered eight of these tags that were relevant to our task and could explain why

Tag	Summ.	Article	Examples
none	0.58	0.77	
Negative	0.21	0.05	counterfeit avert, weep wail
Active	0.26	0.15	intensify overhaul, grasp hop
Strong	0.16	0.03	oust devestate, roar promote
Hostile	0.13	0.01	kidnap ravage, shrug crush
Power	0.08	0.01	curb reclaim, persuade overcome
Positive	0.07	0.03	reinstate mend, reassure hug
Passive	0.05	0.05	deplore mourn, gaze huddle
Weak	0.04	0.03	flounder sag, abandon hesitate

Table 2: Percentage of words in each class covered by different GI tags. The first two example words come from the summary-biased class and last two come from the article-biased class.

a verb has domain-level importance: NEGATIVE, POSITIVE, ACTIVE, PASSIVE, STRONG, WEAK, HOSTILE, and POWER.

Table 2 shows some randomly selected words from each of the eight GI tags. The first two words come from the summary-biased class and last two come from the article-biased class. Table 2 also shows the fraction of verbs in each class that occur in the GI with a given tag, as well as the fraction of verbs that do not have any of the eight tags. It becomes immediately clear that the GI categories do have explanatory power but that it has a major problem with coverage, with the majority of verbs in the summary and article corpora not appearing in the GI at all as shown on the first line. Notably, the coverage is considerably better for the summary-biased verbs. Verbs from several GI categories appeared notably more often in summaries than in articles. For example, verbs with the NEGATIVE tag account for 21% of verbs in summaries, but only 5% of verbs in articles. Other such categories include verbs that imply an active physical engagement (ACTIVE), imply that the actor is in a position of power (STRONG), imply that hostility exists between the entities involved (HOSTILE) or that imply that the actor has the influence to affect the policies of others (POWER). POSITIVE, PASSIVE, and WEAK verbs had more similar appearance rates in both classes, but the absolute number of words covered by these tags was low.

Increasing the strictness of the p-value cutoff for pruning the vocabulary as described in the previous section reduces the size of the vocabulary but increases the purity of the classes by only including verbs that have sufficiently different usage ratios. As shown in Table 3, as we restrict the vocabulary to increasingly certain verbs, the proportion of verbs in the summary class that are tagged by the GI remains almost constant while the proportion of untagged verbs in the article class steadily increases. This indicates that the summary-biased verbs have a consistent distribution of GI tags across all usage ratios, while verbs tended to be tagged less often as the bias towards the articles increased. Although the GI gives good indicators for which words are likely to be important summary words but no indicators for which words are likely to be of no interest in summarizing world news.

p-value	0.01	0.001	0.0001	0.00001
Summary	0.54	0.53	0.54	0.54
Article	0.80	0.86	0.88	0.91

Table 3: Percentage of words with zero GI tags for increasing p-value cutoff strictness

3.1 Personal Stories

To get a sense for what aspects of the verb semantics causes a word to be excluded from the summary, we examined the contexts for the verbs with the highest bias weights in each class. To define the context for each verb, we used the dependency relations produced by the Stanford Parser. Any verb, noun, or adverb placed in a dependency relation with a given verb is considered to co-occur with it. For each of the ten most highly weighted verbs in each class, Table 4 shows the lemmas that co-occurred most frequently with it.

The verbs that are biased towards the articles (not important) seem to capture human element of the news reports, corresponding to passages narrating personal stories of ordinary people involved in the larger political situation discussed in the news. The summary-biased verbs are clearly evocative of the NEGATIVE, ACTIVE, STRONG, HOSTILE, and POWER tags given by the GI and the common usages suggested by their contexts tend to be official, non-personal or that of people in public roles.

No existing resources provide descriptions of this personal vs. non-personal dimension of lexical meaning and we decided to derive such a characterization from data unrelated to the NYT.

Article-biased	
add	country,year,get,States,time,people,do,make
drink	drink,do,take,glass,much,make,eat
sing	song,woman,man,chorus,dance,sing,feel
refer	use,official,attack,part,term,program,day,people
read	time,report,people,write,statement,book,man
think	time,part,get,do,year,take,issue
imagine	take,people,get,come,time,make,ask
remember	day,year,time,see,decade,many
shout	man,people,hear,soldier,get,come,crowd
sit	day,man,road,talk,wall,people,watch,table
Summary-biased	
spur	do,action,help,tell,States,effort,concern,man
hail	leader,man,call,effort,Clinton,step,election,visit
allege	part,case,fraud,arrest,help,people,responsible
avert	attack,Iraq,action,month,confrontation,crisis,official
slay	week,month,member,attack,many,soldier,day,Americans
exile	country,accuse,many,kill,Hussein,family,friend,Arafat
claim	member,bombing,describe,part,group,life,leader
intensify	country,States,war,week,demand,day,year,effort
extradite	Britain,States,try,citizen,Pinochet,trial,member,receive
oust	year,Party,Minister,force,coalition,invasion,month,leader

Table 4: Most frequent co-occurring words for the most extremely weighted verbs.

Personal-biased
threaten wake rain wander kneel yell grin convulse smile chat hug climb gorge nod crouch laugh sleep perch head park
Nonpersonal-biased
acquit deploy misstate founder besiege decriminalize censure peacekeep headquarter streamline dissociate excommunicate unveil deadlock modify extradite ratify imperil chose

Table 5: Top weighted words derived from personal and non-personal blog entries

For this purpose, we used a subset of the ICWSM 2009 Spinn3r Blog Dataset that has been annotated with a semi-supervised classifier trained to identify personal stories (Gordon and Swanson, 2009). We took 56,048 blog entries that had been tagged as being a personal story and 2,196,162 blog entries that were not identified as personal.

We then applied the same procedure that we used for the NYT articles to produce two classes of words: those biased towards blogs describing personal stories and those biased towards non-personal blogs. After restricting the vocabulary to only verbs with a binomial test p-value of at most 0.05, we obtained log odds ratio weights for 3,143 verbs. Of the 1,210 verbs in the NYT classes, 937 were also present in the restricted blog vocabulary. The 20 most and least personal verbs are shown in Table 5.

p-value	GI	GI+blog
0.05	0.134	0.098
0.01	0.130	0.087

Table 6: 10-fold cross-validation mean squared error of a linear regression for increasingly biased vocabularies.

The Pearson correlation between the NYT log odds ratio and the blog log odds ratio is negative and rather high, -0.54, indicating a strong relationship between personal and article-biased words. Restricting the significance to p-value cutoff of 0.01 reduces the vocabulary from 937 to 675 verbs, but strengthens the correlation to -0.61. Of the top 100 summary-biased words, only 18 were personal. Of the top 100 article-biased words, 90 were personal.

Not only do the personal/non-personal classes map on to the summary/article classes well, but they supply explanatory information about words that the GI did not cover. In order to measure this effect, we trained a linear regression to predict the NYT log-odds ratio of a word using a binary feature for each GI tag, as well as a binary feature indicating no tags. We were interested in the reduction of error when the personal-biased information was added. Adding the blog log-odds ratio for each word as a feature improved our results in 10-fold cross-validation, reducing the prediction error by almost 30%. The detailed results are shown in Table 6, for experiments performed for two different p-value cut-offs.

4 Conclusion

We presented a method for data-driven acquisition of domain-level importance of verbs in reports of world news events. Analysis of the acquired verbs reveals that summary-biased words tend to be more negative, active, and hostile, while the article-biased words mostly describe personal actions. This lexicon provides a useful notion of global importance in a domain and can serve as resource for semantic characterization of words in a variety of tasks, including sentence selection in summarization, flagging articles as newsworthy or filtering uninteresting documents. Additionally, we provide a lexicon for personal and non-personal verbs that also captures some of the newsworthiness of the article and summary classes.

References

Jesse Dunietz and Dan Gillick. A new entity salience task with millions of training examples. *EACL 2014*, page 205, 2014.

Andrew Gordon and Reid Swanson. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*, 2009.

Kai Hong and Ani Nenkova. Improving the estimation of word importance for news multi-document summarization. *Proceedings of EACL*, 2014.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.

Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer US, 2012.

Karolina Owczarzak and Hoa Trang Dang. Overview of the tac 2011 summarization track: Guided task and aesop task. In *Proceedings of the Text Analysis Conference (TAC 2011), Gaithersburg, Maryland, USA, November*, 2011.

Dragomir R Radev and Kathleen R McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24 (3):470–500, 1998.

Barry Schiffman, Ani Nenkova, and Kathleen McKeown. Experiments in multidocument summarization. In *Proceedings of the second international conference on Human Language Technology Research*, pages 52–58, 2002.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. 1966.

Kristian Woodsend and Mirella Lapata. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 233–243, 2012.